TSP Table Management Statistical Program for the Analysis of Historical Statistics

Nemeskeri, Istvan; Kovac, Imre

Historical Social Research, Vol. 14 — 1989 — No. 4, 94-98

TSP Table Management Statistical Program for the Analysis of Historical Statistics

*Istvan Nemeskeri, Imre Kovac**

Since the mid-seventies, with the wide spreading of computer techniques, the possibilities and categories of empirical research in the social sciences have changed in a fundamental manner. In Hungary, this change took place somewhat later, in the first half of the 1980s. Explosive change was observable primarily in the case of the processing of sociological data survey, where no research is planned today without reliance on computers. The change is much slower in the case of historical research or research of that nature. Basically, this has two reasons. On the one hand, the statistical methods elaborated for economic and natural phenomena are difficult to adapt to historical problems. On the other hand, the structure of the available data is difficult or impossible to modify and missing data are also difficult if not impossible to recover. So, new methods and programs had to and have to be worked out and the data have to be collected, systematized, unified, recovered, etc. before processing; it is only afterwards that analysis of merit can be carried out.

We have been planning a table management statistical program (TSP) suitable primarily for the analysis of data of an historical nature since 1984 and we have been intensively working on it since 1986.

Simultaneously, we are also planning the establishment of a data bank storing historical data.

Historical Statistics in Hungary

In the course of the work on both the TSP and the planned data bank we rely on the statistics published in Hungary. The TSP shall be developed expressively for the analysis and transformation of statistical materials in a table format. We intend to store primarily the materials of the published sources in the data bank, too.

The historical statistical material published in Hungary can be estimated to fill several hundreds of volumes. Regular statistical surveys have been carried out in Hungary since the last third of the nineteenth century,

*Address all communications to Itsvan Nemeskeri, TARKI, Frankel Leo u. II. H-1027 Budapest, Hungary.

most of which was also published. At the same time, the original manuscripts and statistical sheets of these materials that are of historical value today, have been destroyed due to the events of history (wars, revolutions), hence, the only sources we now have arc the published materials.

The most important historical statistics are the censuses repeated every ten years as from 1870 and the related data surveys; the economic information (data on commerce, transportation, production, prices, land property, factory statistics, animal censuses); health statistics, data on individual social strata (such as the statistical survey of white-collar workers in 1928, or the annual statistics of the colleges and universities (between 1929 and 1933), the statistical surveys of the secondary schools, certain stratum analysis of the Budapest Statistics Office on labourers, traders, merchants, engineers, medical doctors, etc.). These data can be analyzed with the help of TSP and we intend to incorporate these data into the data bank as well. Between 1890 and 1945 statistical surveys were carried out using a similar approach, methods and categories. During this period the formats of the publications and of the tables were also fairly similar. The end of World War II is regarded as the time limit for historical statistics partly because of the change in the spirit and practice of statistics and partly owing to the fundamental change in the economic and social structure of Hungary after 1945. Under the peace treaties following World War I the boundaries of the Hungarian State were significantly modified. The historical statistical materials - in cases of national surveys - naturally applied to the territory of Hungary of the time. Thus, in the period between 1870 and 1944, data were collected from two different territories. As far as the data from the areas regained in the course of World War II arc concerned. Hungarian statist!. cal practice handled these separately (from those obtained from the territory created after World War I).

Naturally, the 1870-1945 time limit does not mean that we have given up the statistics of other historical eras or the unpublished materials in the course of developing the TSP or the construction of the data bank. We regard the processing of the statistic published between 1870 and 1945, turning these into something analyzable and transformable, as the first, although not easy part of our work.

Antecedents

The agricultural data bank started within the organizational framework of the Agricultural Museum can be regarded as the direct antecedent of our plans for the data bank. The work on the agricultural data bank was abandoned despite the numerous results owing to the withdrawal of central funding. These data files shall be accessible at TARK1 from the next year. In the course of the work on the agricultural data bank, two special

programs were prepared: one for drawing maps and the other for the identification of place names that have been changed or differently spelled in the course of time.

The work on the Historical Atlas carried out at the Institute for Historical Science, where computer techniques were used, can be regarded as a methodological antecedent to our work.

TSP program

The present stage of our work entails the preparation of a program using which these statistical data available in table format become easy to handle and to analyze even without specific knowledge of computer techniques. In the course of planning the program, the well-known data management and statistical program packages (such as SPSS, BMDP, DBASE, SAS, etc.) well proved also in the social sciences, have been taken into consideration. It is not our objective to compete with these or to replace these; in fact we intend to ensure adequate link to these program packages: if a task can be solved with these programs, then the data should be easily transferable to these program packages. Researchers dealing with historical statistics, however, have a legitimate claim to carry out the same calculations, estimations or hypothesis analyses with the available statistical data as a sociologist can with the data of the empirical investigations planned by him. Generally, however, these can be carried out using the program packages mentioned above only if the data are available at basic data level. TSP is intended to make up for this deficiency.

The running of the program is supervised by an interactive menu system, in which the user only needs to answer the questions asked and to select from the alternative options. Thus there is no need for special training in the use of the program, yet there is ample opportunity for the tinged formulation of the given problem, for the setting of parameters and for the revision of default values.

The TSP program consists of five main parts:

1. Feeding in, storage and transformation of tables;
2. Statistical calculations and hypothesis examination;
3. Preparation of tables and estimates meeting certain conditions on the basis of known partial information;
4. Examination of Markov-models;
5. Preparation of outputs.

1. When feeding in the data, the user can give the format of the external data stock within a wide range. These data can be stored on a magnetic data carrier accessible to the computer, or, in case of smaller tables, they can be fed in directly from the keyboard. In addition to feeding in the

numerical data, the header of the table and the names of the categories of the individual variables must be defined (just as, e.g. the command VALUE LABELS in SPSS).

The TSP stores the data in its own format, separately the numerical data of the tables and in a separate dictionary the headers describing the tables.

A lot of types of transformations can be carried out with the help of the program:

- merging of categories;
- preparation of marginal tables;
- preparation of conditional tables;
- combination of category variables;
- preparation of percentage tables from frequency tables and vice versa;
- multiplying tables made from samples;
- preparing frequency tables from probability tables and vice versa;
- fitting several tables together;
- breaking up an existing dimension into two or more dimensions.

2. The program calculates the associative measuring indexes that can also be computed with the most frequently used SPSS or BMDP in the case of calculating the statistics of contingency tables. Naturally, the researcher has to decide which of these is adequate to the level of measurement of the table. In its present form, the program contains test of independence conformity and multidimensional analysis of variance of the range of hypothesis test.

3. The main part of the program package is the preparation of new tables which are not included in the original data, of which only partial information is known and where only assumptions can be used. The estimation is carried out using the »iterative proportional fit of the marginal totals« method elaborated for the hierarchic fitting of multidimensional contingency tables, or rather with a modified version of this program. The basic tasks are the following:

a. A and B are known, we want to estimate AxB
b. certain (not necessarily one-dimensional) marginals of a multi-dimensional distribution are known, we which to estimate total distribution;
c. AxC and AxC are known, we wish to estimate BxC;
d. the marginals and certain cells of a contingency table are known, we wish to estimate the table.

These tasks can be formulated if further information is available at a certain sub-set or at an expanded set of the populations under study. For instance, AxB and AxC are known from the national survey and AxBxC is known for all the towns. We wish to estimate BxC for the total population.

Naturally program an attempt is made at estimation also under a few unusual conditions and also at estimating certain extreme distributions, when the target function is not minimized but maximized while meeting the given conditions. Through this it is possible make statements not only like that the frequency of a cell mowed at around a certain value with a high probability, but also, exact, or estimated values can be given for the maximum or minimum of cell frequency.

4. With the help of the program, the fitting of discrete time, discrete status Markov- models (Markov-chains) from the cross-tables of consecutive surveys, the determination of their analytical and statistical characteristics and hypothesis examinations can be carried out. Researchers have great need of this in the case of tables containing certain economic or demographic data.

5. In essence, the program has three different types of output:
   a. the output containing the results of calculations;
   b. the table output of the original and of the estimated cross tables in a format defined by the user, which can be later read and further analyzed by this or some other program;
   c. It is possible to produce a basic data matrix from a multidimensional contingency table (containing original or estimated data) that is suitable for reading in with e.g. SPSS or BMDP and for subjecting it to an optional statistical procedure (naturally, to one that is in line with measurement level and interpretation of the data).

As a first step, the program is prepared for the IBM AT, but, parallel with this, the version for the IBM 3031 mainframe is also under way. The programs for both machines arc written in PASCAL and, as far as possible, the user side of the program is planned and written the same way. According to our plans, the micro-version should be completed in its final form by the end of the year, while the mainframe version should be completed in the first half of 1989.

The storage and handling of the large quantity of tables handled by the program pose a separate task of some difficulty. At the present stage of our work, unfortunately, we do not have the funding needed for its elaboration and realization, so we only have preliminary conceptions about this. Taking the computers presently available in Hungary into consideration, this data base can be envisaged only if the data are stored at different levels. We intend to ensure on-line access only to the verbal information, that is, to the titles and headers of the actual data in a second step.

We hope to be able to continue our research in this direction after the completion of the TSP so that, at the next year's conference, 1 shall be able to inform you of this work and its results.